

## Editorial

**John MacBeath · Lejf Moos**

Received: 11 October 2009 / Accepted: 13 October 2009 /  
Published online: 23 October 2009  
© Springer Science + Business Media, LLC 2009

Evaluation of teaching quality is an issue of international concern and one subject to continuing effort to establish consensus on what constitutes good teaching and how good teaching should be acknowledged and rewarded. In this issue of the journal these theme are addressed from a number of different perspective and country contexts. In turn we visit the United States, Israel, China and Cyprus, in each case reminding us of common tensions and dilemmas but also of how very differently these issues are engaged in differing cultural milieux.

Dilemmas loom large in Yariv's paper. We are here, in the author's own words, in the 'soft' territory of emotions, of embarrassment at having to evaluate teachers' performance, of retracting judgment in response to tearful pleas by teachers. This is not the world of 'hard' performance measures, disciplinary procedures and dismissal but of sensitivity, collegiality and 'keeping it in the family', not 'rocking the boat'. In Yariv's terminology principals preferred to perceive the 'music' behind the words, collecting 'soft' information from hearsay, informal observation and subjective impressions. Student examination performance provided a much less valued source of information than parental, teacher and student complaint.

The qualities valued by principals were primarily personal and moral—motivation, initiative, enthusiasm, honesty, integrity, dependability and loyalty. Laziness, constituted one of the most heinous of sins, threatening 'to tear apart the delicate working relations fabric of the organization'. Sensitivity and love for children were rated highly in what Yariv describes as a child-centred climate in which the greatest offence it to offend a child. Insensitivity was found to be the most disturbing characteristic, the one that principals were least willing to tolerate.

Without an understanding of the cultural and religious context it would be easy to rush to judgement. As Yariv points out, the soft sentiment and the solidarity of the collective reflects a unique Israeli and Jewish cultural value system. For centuries,

---

J. MacBeath · L. Moos (✉)  
Aarhus Univesity, Aarhus, Denmark  
e-mail: moos@dpu.dk

she writes, Jews have lived in exile, been discriminated against and treated as outsiders, their fear of persecution forcing them into their own community life. Their celebration of family as the important unit and children as the center of existence is reflected in their view of the school as an extended family. Observation of a teacher's classroom is perceived as analogous to intruding into a family member's personal space, almost an impolite gesture if you had not been invited.

Nonetheless, poorly performing teachers present one of the toughest challenges school principals have to face, distracting them from other priorities, consuming valuable time which could be given to more growth promoting activities, damaging the school's reputation, diminishing credibility and trust. It is hard to disagree with the author's conclusion that there needs to be more formal strategies in an area fraught with subjective judgment and inconsistency.

The dilemmas, complexities and paradoxes which arise from the casting of students as reliable witnesses to teaching quality is illustrated in Norvilitis and Zhang's paper. While set in a higher education rather than a school context, the researchers findings follow a similar evidential trail. Although stakes are not as high as in the case of China (described in Liu and Teddlie's 2005 study referred to below), the reliability, validity and potential consequences of students' judgments raise worrying questions. Norvilitis and Zhang steer us through this complex maze of reciprocity, leniency, expectation, social comparison effects and the self-serving effect in which students externalize responsibility for negative outcomes. Do students give good evaluation to teachers who are lenient? Do they reward and punish teachers who treat them well or badly? Do their judgments refer to class norms and their own position in relation to these? Do students attribute their poor grades to the fault of the course or the teacher? Do students who receive grades they perceive as fair (that is, what they expect to receive) rate their teachers more positively?

Given well established research findings that students receiving higher grades rate their professors more favorably than their lower achieving counterparts, this study explores that finding through the (ethically contested) strategy of giving students manipulated scores, so testing the above hypothesis once again. Following a mid semester exam students were given their actual earned exam scores and a manipulated class mean that was either 10 percentage points higher or lower than their assigned grade. However, this study neither confirmed the leniency or the reciprocity hypotheses. It appears that students do not simply like classes better when the grading is perceived as easier. Instead, a "selective leniency" may be at work: students like classes that are easy for them, but not classes that are easy for everyone.

The authors do, however, highlight the limitations of the study. It does not address student factors, such as motivation and achievement striving that have, in previous studies, been related to course evaluations. Nor does it address characteristics of teachers such as their age or personality traits such as extraversion. Nor does the study, it is conceded, address a range of extraneous variables which may affect students' attitudes and engagement, assumptions and expectations of their teachers. The gender dimension, is also not accounted for. Norvilitis and Zhang conclude that there remains a need for future research which would explore the impact of these variables on this model.

It is interesting, however, that there is relatively little research on sexuality in the classroom, and one not explored in the above study. This is perhaps because it is highly sensitive and dangerous territory. Yet, as anyone who has taught either in schools or universities is well aware, the interplay of sex and power in the classroom can have a significant impact not only on how the teacher is evaluated but how that latent (and sometimes overt) sexuality may affect the efficacy of teaching and of learning and professional judgment.

We might not be surprised to find such an element missing in Chinese researchers' exploration of what makes a good teacher. As Liu and Meng report, although much has been written about how teachers are evaluated, and often with very high stakes consequences (see Liu and Teddlie 2003, 2005), there have been few empirical studies on the characteristics of good teachers in China. Priority to such research, to what extent might we predict the answers based on similar studies in other countries? How different would we expect Chinese teachers to be from their American counterparts? How significant would the cultural context prove to be?

'In spite of the different categories found in the current study and in the USA, most of the characteristics of good Chinese teachers are consistent with those identified in the USA', conclude the authors, citing studies by Stronge and Hindman (2003) which emphasise characteristics such as classroom management and organization, planning and organizing for instruction, implementing instruction, and monitoring student progress and potential. No surprises there as these are taken-for-granted aspects of what it means to be a classroom teacher. Yet, the more the authors offer parallels the more cultural differences emerge in the subtext.

The headlining of teacher ethics as one of the most important qualities identified by Chinese students is an example of just how much cultural nuance lies beneath the surface. However essential (and from time to time famously observed in the breach) it would not be made so explicit in American studies nor be manifested in the same ways. It would not generally be expressed in behaviour such as using lunch breaks to help students with their studies, nor escorting students to the gate after dismissal and waiting until their parents pick them up.

Where Liu and Meng see Chinese teachers as diverging most from their American counterparts is in relation to test scores, which they write, 'appears to be unique to the Chinese context'. Again such a finding has to be understood in the two respective cultures, as testing in the US is as high stakes as anywhere else in the world, but like ethics, would be implicit rather than an explicit defining 'quality'. The finding that it is students and parents, and not teachers, who are most likely to emphasise test scores is, perhaps much less surprising.

The authors' final conclusion is that 'the difference between the countries in student academic achievement mainly lies in the system of teaching and the cultural roles of teachers in different countries rather than individual teaching beliefs and behaviors'. Such a statement is clearly a matter for debate and needs to be taken in conjunction with Liu and Meng's concluding sentence—'how culture influences teacher effectiveness in different countries certainly merits further exploration'.

Dilemmas continue to figure large in Iasonas Lamprianou's paper. There are persuasive parallels to be drawn between how principals assess their staff and how teachers assess their pupils. Left to their own devices, he writes, teachers will orient

their assessments more towards what it is best to feed back to the child at any point in time but once examination boards raise the assessment stakes, reporting then takes the government, the board or the accountability locus, as its audience, rather than the student. Professionalism is then judged not by formative feedback (see Black and Wiliam in the first volume of this journal) but by ‘consistency of standard’ in external reporting.

‘Teachers are in the business of creating classical unreliability’, writes Lamprianou. Is this because when left to their own judgment rather than tests, interpretation and application of assessment criteria present too complex a task? When teachers are saturated with so much information about their pupils does it not present a particular difficulty in fairly selecting and applying what is salient? How can teachers reliably identify relevant cues from this array of information? In these circumstances the way in which teachers make their judgments “constitutes a prism through which one can examine teachers’ beliefs and values”. So, what teachers do in their varying contexts is to apply different weights to different assessment dimensions according to the cues which they judge as “relevant”.

Lamprianou cites Feiler and Webster’s findings as to teachers’ literacy predictions for an incoming grade one cohort. As he reports their findings, teachers were prepared to make predictions on the most partial of information, gleaned even before the children themselves were encountered in the classroom. Judgments, strongly influenced by perceived social class stereotypes, such as home address or parental occupation, tended to become more permanent over time—the “anchoring-and-adjustment” heuristic. The consequence is that information which confirms initial hypotheses tends to be accorded more weight than evidence which is contradictory, so that selective attention to behaviour which reinforces intuitive first impressions then tends to increase the internal consistency of teachers’ assessments. Albeit inadvertently, predictive validity becomes enhanced.

On the student side, when evaluation is normative, highly differentiated and with emphasis on social comparison, students in high performing classes underestimate their own performance through social comparison while those at the other end internalise their low status—‘shades of the self-fulfilling prophecy’.

These issues are discussed within the overarching theme of SBA, or school-based assessment which contains many dilemmas and tensions of its own. Lamprianou agrees with Paul Black that ‘it is merely a “dream”’ to imagine that requiring teachers to conduct SBA without adequate in-service support will not of itself embed formative assessment in everyday classroom practice. ‘Our data set suggests’ concludes the author, that what does get embedded is a reporting strategy. To paraphrase an old nostrum—What gets attention gets embedded.

If we needed any further evidence that assessment is an emotional affair then Lipnevich and Smith’s paper provides it. What is perhaps surprising, but ought not to be, is the degree to which students’ sense of self worth is affected by the judgments that authority figures, such as professors, make about their work.

What also comes through strongly from the authors’ review of previous research is the continuing ambiguity as to the kind of feedback most likely to promote students’ progress and engagement with the task at hand and minimize the emotional impact of the feedback they are given. Do grades motivate or demotivate? What is the effect of praise, comment, or specific advice? What is the ideal balance between

summative and formative feedback? It is the inter-relationship among these that generates the most disagreement among researchers, especially when most studies are conducted in a context where marks and grades are deeply embedded in the educational psyche and are attended by so much emotional baggage.

Lipnevich and Smith's experimental study was designed to determine which permutation of grades, praise, formative comment was seen, from the student's viewpoint, as most and least helpful. To what extent was the source of the feedback critical — the authority figure of the professor or the disinterested computer programme? The experimental design with each of eight different students groups receiving different combinations of feedback and from (allegedly) different human and inanimate sources led to some clear conclusions.

Feedback from a neutral source such as a computer defused some of the emotional Angst, embarrassment and even anger of being given bad news by a professor. Fragile self esteem could take a knock when a professor's comments were taken as a personal slight. However, students tended to feel that computer-provided suggestions for improvement were faulty and irrelevant even though the source of these same comments were, in another experimental group, given as coming from the professor.

Students were unanimous in their view that grades were ineffective when mastery of learning was needed, although typically a grade is a course, institutional or political imperative. There was a high degree of unanimity that detailed comments were the most effective form of feedback. The use of praise by the teacher remains a contested issue as research findings are that praise can both impede and facilitate students' performance, one argument being that praise directed at the self robs cognitive resources that would otherwise be committed to the task. On the positive side praise could provide a 'buffer' or 'sugar coating' when something untoward was coming your way, cushioning the impact of bad news. While low marks tended generally to have a depressing effect on motivation, high marks tended to be a disincentive to improve.

Although this is not a themed issue of the journal the consonance of issues raised in these papers is striking. We are reminded in every paper that assessment, evaluation and accountability are not simply technical considerations but that they have to do with human beings, human sensitivities and human motivation. As we have learned, it is not just small children whose egos are deeply involved in their academic endeavours but grown up university students and mature teachers. Perhaps there is a message here for policy makers and politicians who, while legislating for an emotion free system, are themselves highly vulnerable to adverse feedback and invest considerable emotional energy in looking out for their ratings. There is much in these papers for their enlightenment.